大数据时代的机器学习热点

国际机器学习大会ICML 2013参会感想

文 / 王威廉

国际机器学习大会（ICML）源于1980年在卡内基-梅隆大学（CMU）举办的机器学习研讨会。几十年过去了，ICML如今已发展为由国际机器学习学会（IMLS）主办的年度机器学习国际顶级会议，可以说代表了当今机器学习学术界的最高水平。那么，在“大数据”时代的背景下，ICML又有什么看点呢？今年，第三十届国际机器学习大会（ICML 2013）于6月16-21日在美国亚特兰大举行，下面我与各位读者一起分享一下我的参会感想。

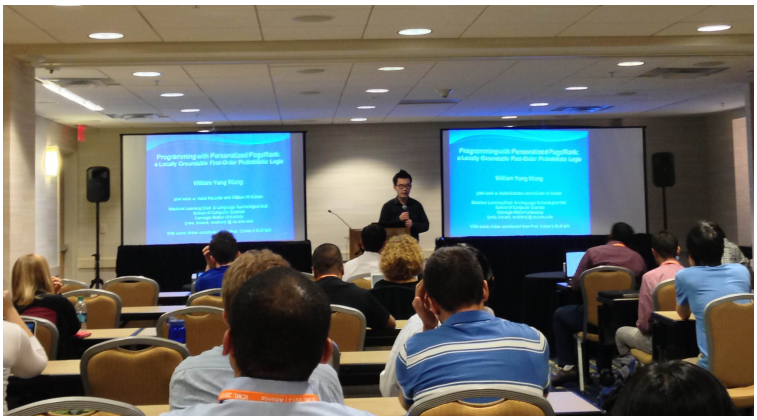
可扩展的大规模图学习与推断算法

可扩展性（Scalability）可谓是贯穿今年ICML的一大主线。首先，什么是可扩展性？通俗的说，就是让传统的机器学习算法能够适应并处理海量数据（如上百亿级别的文件）。在结构化数据普遍存在的今天，可扩展的图算法，尤其是可扩展的复杂概率图算法尤其引人注目。到底实现可扩展的图结构算法有什么困难？一个显而易见的难点就在于：数据样本之间往往有较强的依赖性，所以MapReduce这种对数据进行“分割-计算-合并”处理的传统数据并行化方法可能并不直接适用于图结构的并行化。

在ICML开幕前一天的结构化学习研讨会上，Facebook数据科学家Jonathan Chang就介绍了他们面临的实际问题：Facebook的在线社交网络有大约109个结点（用户），以及大约1012条边（关系）。在这种规模的图结构里，就算仅仅是计算所有用户好友的好友（Friends of Friends）这一简单属性，如果不使用高效的图计算模型，也可能产生庞大的开销和非最优的结果。Jonathan接着介绍了他们的解决方法：Giraph，一种基于图灵奖得主Leslie Valiant在20世纪80年代推出的Bulk Synchronous Parallel（BSP）模型衍生而来的开源工具。Giraph其实可以被看成是近年来Google Pregel迭代计算模型的开源版本：在这个以结点为中心的模型的每次迭代计算中，结点处理上次收到的消息，发送消息给其他结点，并且改变自身结点、边或者拓扑结构。当Jonathan被问到与GraphLab的对比时，他表示Facebook曾经尝试过GraphLab，但并不能达到他们的需求。非常有意思的是，Carlos Guestrin正好将在本次ICML大会上做关于GraphLab最新进展的主题报告。

第二天一早，会场早已座无虚席。在ICML大会主席Michael Littman简短介绍后，Carlos Guestrin，这位机器学习的新生代领军人物就正式登台了。说到GraphLab，相信大家不会过于陌生：GraphLab三年前诞生于CMU机器学习系，主要目的是为了并行化复杂的图算法。

Carlos接下来介绍了他们开发GraphLab的心路历程：早期推出的第一代GraphLab，在许多任务中取得了非常惊人的表现。如GraphLab1对于Never-Ending Language Learning（Tom Mitchell的永不停息机器学习系统）的CoEM算法的并行化实验，所需时间仅仅是Hadoop的0.3%。然而，GraphLab1在处理14亿结点、67亿条边的Altavista数据集上失败了。为什么呢？在分析了数据后，他们发现Altavista服从自然图的Power Law分布属性，即有



作者在ICML研讨会上介绍概率编程的最新进展（拍摄者：Cheng Zhang）

Andrew Ng在ICML迁移学习研讨会上做关于深度学习的远程演讲

1%的结点与53%的边相连，而这些高度数的结点会导致他们原先的算法失效，并且使得图结构很难被分割。同时，他也介绍了Pregel的问题，由于Pregel/Giraph是同步类算法，很多情况效率也不如非同步算法，在自然图上也会发生此类问题。2012年推出的GraphLab2对自然图计算的瓶颈问题进行了改进：通过把计算迁移到数据上，他们设法并行化高度数的结点，并且设计了有效的适应自然图Power Law分布的图分割算法。如今，GraphLab2在处理Altavista的数据上已经有了重大突破，使用1024个核与4.4TB的内存，现在只需要11分钟的处理时间。最后，Carlos介绍了GraphLab3的规划：GraphLab3将结合第一代的代码可读性与第二代强大的可扩展性特点，使得图并行算法能被更多的开发者所使用。另外值得注意的是，如今GraphLab已正式注册了公司，并且获得了675万美元的风险投资。

深度学习热潮的延续

随着深度学习概念的兴起，本届ICML自然也是少不了许多关于特征学习以及深度神经网络的工作。由于深度学习的学术界领头人Geoffrey Hinton老先生已归顺了Google，所以加拿大蒙特利尔大学的Yoshua Bengio教授在本次大会中显得非常活跃。首先在6月16日的研讨会上，Yoshua介绍了他近期一些较为“激进”的思想：他认为传统的隐变量概率图模型在实际使用中会产生很多的局部最优区域，这些局部最优区域甚至可能会超过经典马尔科夫链蒙特卡洛（MCMC）推断算法的采样次数，最终导致得到非优的推断结果。Yoshua提出，传统的隐变量模型可以被Denoising Autoencoders（DA）替代。DA可以被看作是一种生成式深度学习模型（generative model），并可使用任意的变量（离散或连续）、任意的噪音，以及任意的损失函数。Yoshua最新研究成果表明，DA不仅在输入层，在中间计算层也可以加入噪音建模。他认为此算法可以用经典的反向传播算法训练参数，从而克服显式传统隐变量模型的缺点。在6月17日的大会上，Yoshua还有一项有意思的工作就是介绍Recurrent Neural Networks训练过程中梯度（gradient）的消失与爆炸（过大）现象。其实梯度的突然消失与爆炸在各类随机梯度下降算法中普遍存在，也是一个优化中常见的问题。他们解决的方法是将爆炸的梯度重新规整，并且将消失的梯度正则化。

6月19日，Google语音搜索组Vincent Vanhoucke做了关于深度学习在语音识别中应用的精彩主题演讲。Vincent从语音的基础（声学模型与语言模型），堪称经典的高斯混合模型-隐马尔科夫模型，语者适应技术，讲到如今基于深度学习的语音识别。深度学习在语音学习的应用源自一个跨领域的经典合作：故事是2010年前后，微软和Google的语音组分别招了Hinton老先生的几个学生做实习，结果发现如果不用传统的MFCC/PLP特征，而用深度学习直接从语音信号里学习特征，并且用深度学习技术对声学模型建模，居然可以在标准数据集TIMIT上取得惊人的突破。以Google为例，3个月时

间下来，语音搜索的相对错误率竟然减少了10%。Vincent介绍说，其实语音识别对神经网络并不陌生，早在20世纪80年代末与90年代，神经网络就在语音及音素识别上有了应用，但基于当时算法和硬件的限制，并没有被广泛采纳。随后神经网络在语音世界里消失了近10年，直到2010年前后的深度学习热潮，才重新回到人们的视野里。

还有一个不得不提的就是斯坦福大学Andrew Ng关于用GPU做深度学习的最新工作。还记得Google曾经用1000台计算机（开销约100万美元）做的猫脸识别软件吗？在本次ICML中，Andrew的学生仅用价值2万美元的GPU集群，就做到了相同的准确率。可以说，Andrew的这项GPU技术，使得深度学习技术逐步走向中小公司及学校，又迈进了一大步。同时，在6月21日的迁移学习研讨班中，Andrew还通过Skype视频远程与我们进行了沟通，介绍了斯坦福大学深度学习项目的研究进展，尤其是在计算机视觉上的应用。另外，在ICML的讲习班里，另一位深度学习的领路人，纽约大学的Yann LeCun教授也做了一个长达3小时的深度学习教学讲座，受到了各位听众的好评。

其他机器学习热点问题及最新进展

本年度ICML的经典论文奖颁给了10年前（ICML 2003）两篇来自CMU的论文：第一篇论文是Jerry Zhu、Zoubin Ghahramani以及John Lafferty关于图结构半监督学习的经典论文。如果你关心机器学习的进展，不难发现，用半监督或无监督学习方法挖掘无标签的数据，不仅是过去10年，还很可能是大数据时代的一个热点。另外一篇是Martin Zinkevich的在线学习经典论文。在线学习解决的问题是：当数据集太大，并且数据流速度太快的情况下，我们没有理由每次都把所有数据全部重新训练一遍。通过在线学习方法，我们可以不用把数据存在硬盘里，每次直接用实时的数据流来更新机器学习模型的参数。另外，ICML 2013最佳论文奖之一授予了Vanishing Component Analysis。传统的特征选择方法通常是在采样中选择显著的特征，这篇论文研究的是，在特征选择时，能不能选择一些不变的特征呢？在特征选择的问题中，这也是一个比较新的研究方向。

如果你是Dave Blei的粉丝或者对文本分析有兴趣，ICML 2013也有相当多有意思的主题建模文章，如Arora等人推出的基于锚点词（anchor words）的主题建模新算法，Ke Zhai等人的无限词汇维度在线LDA模型，以及Weicong Ding等人推出的基于投影方法的主题模型都让人眼前一亮。

核函数领域的专家Alex Smola在ICML上介绍了一种名为Fastfood的核函数计算方法，使得计算核函数的时间和空间复杂度分别降到了O(nlogd)与O(n)。这对广大的基于非线性核函数的SVM应用来讲，绝对是一个大救星。

最后还有就是概率编程（probabilistic programming）：虽然本次大会关于概率编程的研究不多，但其日前被DARPA认为是机器学习的未来。概率编程的主要思想就是对确定性编程语言概率化，使得不具备机器学习专业背景的程序员也可以用简单的程序语言与规则来从数据中学习规律，对未知世界进行预测。IMLS主席William Cohen教授与我分别在16日与20日的研讨班上简单介绍了新发明的高效概率化Prolog语言ProPPR：通过几行简单的逻辑编程，可以在复杂的图结构上进行快速的推断，并且实现统计关系推断、分类、实体消歧、序列预测等多种任务。

通过本次大会，我们不难发现，随着大数据时代的来临，机器学习领域也正在悄然积极应对。值得一提的是，ICML 2014将于明年的6月21-26日在中国北京举行，届时中国的机器学习爱好者将有机会在家门口享受一场机器学习的饕餮盛宴。

王威廉

毕业于哥伦比亚大学，目前在CMU攻读博士。曾供职于微软总部研究院、哥大工学院、南加州大学。ACL、CIKM、COLING、Interspeech等知名国际会议上发表论文20余篇，并担任多家SCI杂志的审稿人。2011年被CMU校长选为R. K. Mellon Presidential Fellow。